# Technical Supplement to *How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies*

Jerry Brunner

September 12, 2016

**Abstract**

This document provides technical details and derivations. In order to make the exposition self-contained, some material from the main paper is repeated. Just a few references are given. For standard material from mathematical statistics, consult the index of a text such as Stuart and Ord (1999).

When a statistical test is applied to data and the results are reported in the scientific literature, the power of the test evaluated at the true parameter values (including the case where the null hypothesis might be true) is a single unknown quantity. This is what we mean by the "true power" of the test. Sometimes we will just call it "power."

Our objective is to estimate the mean true power of a population of tests that rejected the null hypothesis. In most cases, the tests will vary in the degree to which the null hypothesis is wrong (effect size), as well as sample size and other aspects of the study design that determine power. We call this "heterogeneity."

Assume an upper-tailed test, so that the null hypothesis will be rejected at significance level $\alpha$ when some continuous test statistic $T$ exceeds a critical value $c$. The cumulative distribution function, inverse cumulative distribution and density of a test statistic $T$ are written $\texttt{p}(t,\texttt{ncp})$, $\texttt{q}(t,\texttt{ncp})$ and $\texttt{d}(t,\texttt{ncp})$ respectively. The notation suggests that the distribution of $T$ is one of the common non-central distributions like the non-central $t$, $F$ or chi-square. In fact it is more general, and will refer to the distribution of any test statistic under any specific alternative except when special properties of the non-central distributions are used. If the null hypothesis is true (or when the null hypothesis involves inequalities and the true parameters are on the boundary of the region specified by the null hypothesis) with $\texttt{ncp} = 0$, we will write $\texttt{p}(t)$, $\texttt{q}(t)$ and $\texttt{d}(t)$.

# 1 Non-central distributions

For the non-central $Z$, $t$, chi-squared and $F$ distributions, the non-centrality parameter can be factored into the product of two terms. The first term is a function of sample size, and the second term is a function of the unknown parameters that reflects how wrong the null hypothesis is. In symbols,

$$\texttt{ncp} = f_1(n)\, f_2(\texttt{es}), \tag{1.1}$$

where $n$ is the sample size and $\texttt{es}$ is *effect size.* As we use the term, effect size refers to any function of the model parameters that equals zero when the null hypothesis becomes false, and assumes larger and larger positive values as the null hypothesis becomes more false. In Equation (1.1), $f_1(n)$ and $f_2(\texttt{es})$ are strictly increasing functions. Here are some examples.

## 1.1 $Z$ and $t$-tests

Consider a random sample of $n$ observations from a population with mean $\mu$ and standard deviation $\sigma$. For testing the null hypothesis $H_0 : \mu \leq \mu_0$ against the alternative $H_1 : \mu > \mu_0$, the usual test statistic is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

where $S$ is the sample standard deviation. First, make it a $Z$-test. The observations may not be normally distributed, but assume that the sample is large enough so that the sampling distribution of $\overline{X}$ is approximately normal, and that $S$ may be substituted for $\sigma$ without much loss of accuracy.

If the null hypothesis is true with $\mu = \mu_0$, then subtracting $\mu_0$ from $\overline{X}$ "centers" the sampling distribution, making its mean equal to zero. In this case the distribution of $T$ is approximately standard normal. If $\mu \neq \mu_0$ the distribution of $T$ is non-central, and

$$\begin{aligned}
T &= \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{\overline{X} - \mu + \mu - \mu_0}{\sigma/\sqrt{n}} \\
&= \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) + \left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \\
&= Z + \sqrt{n}\left(\frac{\mu - \mu_0}{\sigma}\right) \\
&= Z + \texttt{ncp},
\end{aligned}$$

where $Z$ is standard normal and the non-centrality parameter $\texttt{ncp} = \sqrt{n}\left(\frac{\mu-\mu_0}{\sigma}\right)$. Thus the sampling distribution of $T$ is approximately normal, with mean $\texttt{ncp}$ and standard deviation one. The non-centrality parameter may also be written $\sqrt{n} \cdot \mathbf{d}$, where $\mathbf{d}$ is Cohen's (1988, p. 46) effect size measure $\mathbf{d}$. This fits Formula (1.1), with $f_1(n) = \sqrt{n}$ and $f_2(\texttt{es}) = \texttt{es}$.

Similar calculations apply to tests based on any statistic with a sampling distribution that is approximately normal for large samples. This covers important cases like the $Z$-tests for regression coefficients in logistic regression and the $Z$-tests for the parameters of structural equation models. And if the data in the example above are normally distributed, `ncp` is exactly the non-centrality parameter of a non-central $t$ distribution.

## 1.2 $F$-tests

In the preceding discussion of the $Z$ and $t$-tests, note that the non-centrality parameter has the same form as the test statistic, but with Greek-letter parameters substituted for the corresponding Roman-letter statistics. This pattern extends to the general linear model with fixed effects. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$\mathbf{X}$ is an $n \times p$ matrix of known constants

The columns of $\mathbf{X}$ are linearly independent

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants

$\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n)$

$\sigma^2 > 0$ is an unknown constant.

The general linear test of $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ is based on the test statistic

$$F^* = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})/q}{MSE}$$

where the $q$ rows of $\mathbf{L}$ are linearly independent, so that $q \leq p$. Under the null hypothesis, the distribution of $T$ is central $F$ with numerator degrees of freedom $q$ and denominator degrees of freedom $n-p$. When $H_0$ is false, the test statistic $F^*$ has a non-central $F$ distribution with degrees of freedom $q$ and $n-p$, and non-centrality parameter

$$\texttt{ncp} = \frac{(\mathbf{L}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\boldsymbol{\beta} - \mathbf{h})}{\sigma^2}. \qquad (1.2)$$

Note the similarity of the non-centrality parameter to the test statistic. It is almost irresistible to estimate `ncp` with $(q-1)F^*$, which is the basis of the failed observed power method.

To see that the expression for `ncp` in (1.2) actually has the multiplicative form (1.1), multiply and divide by $n$ to obtain.

$$\texttt{ncp} = n\,\frac{(\mathbf{L}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{L}(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\boldsymbol{\beta} - \mathbf{h})}{\sigma^2}.$$

This is of the form (1.1), with $f_1(n) = n$.

Division of $\mathbf{X}'\mathbf{X}$ by $n$ is not arbitrary. It converts the SSCP matrix to a matrix of moments that is constant for experimental designs with designated proportions of observations allocated to the various treatments. For more general regression models, it settles down rapidly to a constant matrix as $n$ increases, by the Law of Large Numbers.

Now specialize to the case of a one-way analysis of variance with $k$ treatments and equal sample sizes. The formula for the $F$ statistic is

$$F^* = n\,\frac{\frac{1}{k}\sum_{j=1}^{k}(\overline{Y}_j - \overline{Y})^2}{(p-1)MSE},$$

where $n$ is the total sample size and $\overline{Y}$ is the arithmetic mean of the sample treatment means $\overline{Y}_j$. The non-centrality parameter is

$$\texttt{ncp} = n\,\frac{\frac{1}{k}\sum_{j=1}^{k}(\mu_j - \overline{\mu})^2}{\sigma^2},$$

where $\overline{\mu}$ is the arithmetic mean of the population treatment means $\mu_j$. Correspondence with expression (1.1) is apparent. It is quite natural to say that the non-centrality parameter is the sample size multiplied by "effect size," a kind of variance ratio or signal-to-noise ratio. Cohen's (1988, p. 275) effect size $\mathbf{f}$ for this problem is the square root of the variance ratio, so that

$$\texttt{ncp} = f_1(n)\, f_2(\texttt{es}) = n\,\texttt{es}^2.$$

As an example of how Formula (1.1) can accomodate different definitions of effect size, suppose we adopt Cohen's (1988, p.281) alternative effect size $\eta = \sqrt{\frac{\mathbf{f}^2}{1+\mathbf{f}^2}}$ for this same problem. Then

$$\texttt{ncp} = f_1(n)\, f_2(\texttt{es}) = n\,\frac{\texttt{es}^2}{1-\texttt{es}^2}.$$

## 1.3 $\chi^2$-tests

Tests statistics and non-centrality parameters also take very similar forms for commonly used chi-squared tests. In an $a \times b$ contingency table with $n$ observations, let $p_{ij}$ denote the sample proportion in the cell from row $i$ and column $j$, while $\pi_{ij}$ is the corresponding true probability. The marginal totals are

$$p_{i.} = \textstyle\sum_{j=1}^{b} p_{ij} \quad p_{.j} = \sum_{1=1}^{a} p_{ij} \quad \pi_{i.} = \sum_{j=1}^{b} \pi_{ij} \quad \pi_{.j} = \sum_{1=1}^{a} \pi_{ij}.$$

Then the test statistic for the the common Pearson chi-squared test of independence may be written

$$T = n\sum_{1=1}^{a}\sum_{j=1}^{b}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}},$$

and the non-centrality parameter is

$$\texttt{ncp} = n\sum_{1=1}^{a}\sum_{j=1}^{b}\frac{(\pi_{ij} - \pi_{i.}\pi_{.j})^2}{\pi_{i.}\pi_{.j}} \tag{1.3}$$

(Agresti 1990, p. 241). Taking effect size to be Cohen's (1988, p. 216) **w**, Expression (1.3) becomes

$$\texttt{ncp} = f_1(n)\, f_2(\texttt{es}) = n\, \texttt{es}^2.$$

# 2 Two Populations of Power

When statistical tests with varying power are randomly sampled from some population, power becomes a random quantity with its own probability distribution. This section establishes some principles relating the distribution of power before selection for significance to its distribution after selection. The principles are illustrated with a numerical example. The example is based on a population of $F$-tests with 3 and 26 degrees of freedom, with varying power. Variation in power comes from variation in the non-centrality parameter, which is sampled from a chi-squared distribution with degrees of freedom chosen so that population mean power is very close to 0.80.

Denoting a randomly selected power value by $G$ and the non-centrality parameter by $\lambda$, population mean power is

$$E(G) = \int_0^\infty (1 - \texttt{pf}(c, \texttt{ncp} = \lambda))\, \texttt{dchisq}(\lambda)\, d\lambda$$

To verify the numerical value of expected power for the example,

```
> alpha = 0.05; criticalvalue = qf(1-alpha,3,26)
> fun = function(ncp,DF) (1 - pf(criticalvalue,df1=3,df2=26,ncp))*dchisq(ncp,DF)
> integrate(fun,0,Inf,DF=14.36826) # Strange fractional df were located using uniroot
0.8000001 with absolute error < 5.9e-06
```

As the comment statement says, the "strange fractional `df` were located using `uniroot`." Specifically, the absolute difference between the output of `integrate` and the value 0.8 was minimized numerically over the degrees of freedom value. The minimum occurred at 14.36826.

**Principle 1** *Population mean power equals the probability of a significant result.*

**Proof** Suppose that the distribution of $G$ is discrete. The probability of rejecting the null hypothesis is

$$\begin{aligned} Pr\{T > c\} &= \sum_g Pr\{T > c | G = g\} Pr\{G = g\} \\ &= \sum_g g\, Pr\{G = g\} \\ &= E(G), \end{aligned} \tag{2.1}$$

which is population mean power. If the distribution of power is continuous with probability density function $f_{_G}(g)$, the calculation is

$$\begin{aligned} Pr\{T > c\} &= \int_0^1 Pr\{T > c|G = g\} f_{_G}(g)\, dg \\ &= \int_0^1 g\, f_{_G}(g)\, dg \\ &= E(G) \quad \blacksquare \end{aligned}$$

Continuing with the numerical example, we first sample one million non-centrality parameter values from the chi-squared distribution that yields an expected power of 80%. These values are in the vector `NCP`. We then calculate the corresponding power values, placing them in the vector `Power`. Next, we generate one million random $F$ statistics from non-central $F$ distributions, using the non-centrality parameter values in `NCP`. In the R output below, observe that mean power is very close to the proportion of $F$ statistics exceeding the critical value. This illustrates Principle 1.

```
> popsize = 1000000; set.seed(9999)
> NCP = rchisq(popsize,df=14.36826)
> Power = 1 - pf(criticalvalue,df1=3,df2=26,NCP)
> mean(Power)
[1] 0.8002137
> Fstat = rf(popsize,df1=3,df2=26,NCP)
> sigF = subset(Fstat,Fstat>criticalvalue)
> length(sigF)/popsize # Proportion significant
[1] 0.800177
```

The sub-population of power values for tests that are statistically significant is of great interest, because the mean of this sub-population is exactly the probability of replicating a randomly selected finding. To see this, think of a coin-tossing experiment in which a large population of coins is manufactured, each with a different probability of heads. All the coins are tossed, and only the ones showing heads are retained. One of these is randomly selected, and tossed again (exact replication). By Principle 1, the probability of observing a head is exactly the mean probability of a head for the sub-population of coins that were retained.

Continuing the numerical example, the sub-population of power values corresponding to significant results are stored in `SigPower`. The tests that were significant are repeated (with the same non-centrality parameters), and the test statistics placed in `Fstat2`. The proportion of test statistics in `Fstat2` that are significant is very close to the mean of `SigPower`. This gives empirical support to the statement that population mean power after selection for significance equals the probability of obtaining a significant result again.

```
> SigPower = subset(Power,Fstat>criticalvalue)
> mean(SigPower) # Population mean power after selection for significance
[1] 0.8274357
> # Replicate the tests that were significant.
> sigNCP = subset(NCP,Fstat>criticalvalue)
> Fstat2 = rf(length(sigF),df1=3,df2=26,ncp=sigNCP)
> # Proportion of replications significant
> length(subset(Fstat2,Fstat2>criticalvalue)) / length(sigF)
[1] 0.827172
```

**Principle 2** *The effect of selection for significance is to multiply the probability of each power value by a quantity equal to the power value itself, divided by population mean power before selection. If the distribution of power is continuous, this statement applies to the value of the probability density function.*

**Proof** Suppose the distribution of power is discrete. Using Bayes' Theorem,

$$Pr\{G = g|T > c\} = \frac{Pr\{T > c|G = g\}Pr\{G = g\}}{Pr\{T > c\}} = \frac{g\,Pr\{G = g\}}{E(G)}. \tag{2.2}$$

If the distribution of power is continuous with density $f_{_G}(g)$,

$$\begin{aligned} Pr\{G \leq g|T > c\} &= \frac{Pr\{G \leq g, T > c\}}{Pr\{T > c\}} \\ &= \frac{\int_0^g Pr\{T > c|G = x\} f_{_G}(x)\,dx}{E(G)} \\ &= \frac{\int_0^g x\, f_{_G}(x)\,dx}{E(G)}. \end{aligned}$$

By the Fundamental Theorem of Calculus, the conditional density of power given significance is

$$\frac{d}{dg} Pr\{G \leq g|T > c\} = \frac{g\,f_{_G}(g)}{E(G)}. \quad \blacksquare \tag{2.3}$$

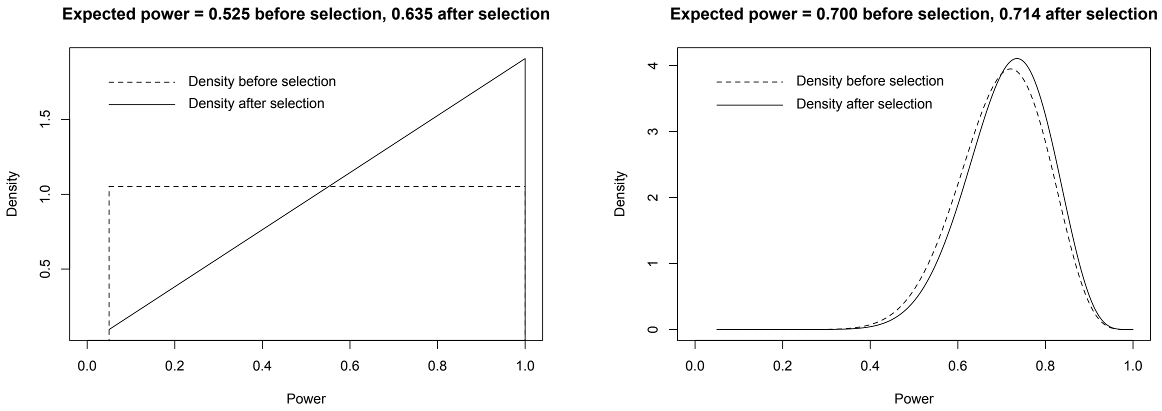
Figure 1 contains two illustrations of Principle 2. The left panel is a simple, artificial example in which power before selection is uniformly distributed on the interval from 0.05 to 1.0, with an average power of 0.525. The corresponding distribution after selection for significance is triangular and with a mean power of 0.635, a substantial change. In the right panel of Figure 1, power before selection is less heterogeneous, and higher on average than in the first example. Average power before selection is .70. After selection the distribution is changed only slightly, with an average power of 0.714.

**Principle 3** *Population mean power after selection for significance equals the population mean of squared power before selection, divided by the population mean of power before selection.*

**Proof** Suppose that the distribution of power is discrete. Then using (2.2),

$$E(G|T > c) = \sum_g g\,\frac{g\,Pr\{G = g\}}{E(G)} = \frac{E(G^2)}{E(G)}. \tag{2.4}$$

Figure 1: Population distribution of power before and after selection for significance


Expected power = 0.525 before selection, 0.635 after selection

Expected power = 0.700 before selection, 0.714 after selection

If the distribution of power is continuous, (2.3) is used to obtain

$$E(G|T > c) = \int_0^1 g \, \frac{g \, f_G(g)}{E(G)} \, dg = \frac{E(G^2)}{E(G)}. \quad \blacksquare \tag{2.5}$$

In the example, `SigPower` contains the sub-population of power values corresponding to significant results. Observe the verification of Formula 2.5.

```
> # Repeating ...
> SigPower = subset(Power,Fstat>criticalvalue)
> mean(SigPower)
[1] 0.8274357
> mean(Power^2)/mean(Power)
[1] 0.8275373
```

It is also possible to recover the original probability of a significant result from the distribution of power given significance.

**Principle 4** *Population mean power before selection equals one divided by the population mean of the reciprocal of power after selection.*

**Proof** Using Formula 2.2,

$$\begin{aligned}
E\left(\frac{1}{G}\middle| T > c\right) &= \sum_g \left(\frac{1}{g}\right) \frac{g \, Pr\{G = g\}}{E(G)} \\
&= \frac{1}{E(G)} \sum_g Pr\{G = g\} = \frac{1}{E(G)} \cdot 1 \\
&= \frac{1}{E(G)},
\end{aligned}$$

so that

$$E(G) = 1 \Bigg/ E\left(\frac{1}{G}\middle| T > c\right).$$

A similar calculation applies in the continuous case. ■

Principle 4 is illustrated below.

```
> 1/mean(1/SigPower)
[1] 0.8000502
```

In the example, population mean power is 0.80, while population mean power given significance is roughly 0.83. It is reasonable that selecting significant tests would also tend to select higher power values on average, and in fact this intuition is correct. Since

$$\begin{array}{rcl} Var(G) & = & E(G^2) - (E(G))^2 \geq 0, \text{ we have} \\ E(G^2) & \geq & (E(G))^2, \text{ and hence} \\ \dfrac{E(G^2)}{E(G)} & \geq & E(G). \end{array}$$

Principle 3 says $\frac{E(G^2)}{E(G)} = E(G|T > c)$, so that $E(G|T > c) \geq EG)$. That is, population mean power given significance is greater than the mean power of the entire population, except in the homogeneous case where $Var(G) = 0$. The exact amount of increase has a compact and somewhat surprising form.

**Principle 5** *The increase in population mean power due to selection for significance equals the population variance of power before selection divided by the population mean of power before selection.*

### Proof

$$\begin{array}{rcl} E(G|T > c) - E(G) & = & \dfrac{E(G^2)}{E(G)} - E(G) \\ & = & \dfrac{E(G^2)}{E(G)} - \dfrac{(E(G))^2}{E(G)} \\ & = & \dfrac{Var(G)}{E(G)}. \quad ■ \end{array}$$

Illustrating Principle 5 for the ongoing example,

```
> mean(SigPower) - mean(Power)
[1] 0.02722205
> var(Power)/mean(Power)
[1] 0.02732371
```

# 3 Maximum likelihood

Even though sample size is a random variable, the quantities $n_1, \ldots, n_k$ are treated as fixed constants. This is similar to the way that $x$ values in regression and logistic regression are treated as fixed constants in the development of the theory, even though clearly they are often random variables. Making the estimation conditional on the observed values $n_1, \ldots, n_k$ allows it to be distribution free with respect to sample size, just as regression and logistic regression are distribution free with respect to $x$. This is much better than adopting parametric assumptions about the joint distribution of sample size and effect size.

The likelihood function given significance is a product of conditional densities. When there is a single fixed effect size, that effect size is the only unknown parameter. In this case, the conditional density of the test statistic $T_j$ given significance and given the random sample size $N_j$ being equal to the constant $n_j$ is

$$\begin{aligned}
\frac{d}{dt}P\{T_j \leq t | T_j > c_j, N_j = n_j\} &= \frac{d}{dt}\frac{P\{T_j \leq t, T_j > c_j, N_j = n_j\}}{P\{T_j > c_j, N_j = n_j\}} \\
&= \frac{d}{dt}\frac{P\{c_j < T_j \leq t, N_j = n_j\}}{P\{T_j > c_j, N_j = n_j\}} \\
&= \frac{d}{dt}\frac{P\{c_j < T_j \leq t | N_j = n_j\}P\{N_j = n_j\}}{P\{T_j > c_j | N_j = n_j\}P\{N_j = n_j\}} \\
&= \frac{d}{dt}\frac{\texttt{p}(t, f_1(n_j)f_2(\texttt{es})) - \texttt{p}(c_j, f_1(n_j)f_2(\texttt{es}))}{1 - \texttt{p}(c_j, f_1(n_j)f_2(\texttt{es}))} \\
&= \frac{\frac{d}{dt}\texttt{p}(t, f_1(n_j)f_2(\texttt{es}))}{1 - \texttt{p}(c_j, f_1(n_j)f_2(\texttt{es}))} \\
&= \frac{\texttt{d}(t, f_1(n_j)f_2(\texttt{es}))}{1 - \texttt{p}(c_j, f_1(n_j)f_2(\texttt{es}))},
\end{aligned}$$

yielding Expression (3.1) in the main paper.

Suppose there is heterogeneity in both sample size and effect size, and that effect size is continuous. The joint probability distribution of sample size and effect size before selection is determined by the marginal distribution of sample size $P\{N = n\}$ and the conditional density of effect size given sample size $g_\theta(\texttt{es}|n)$, where $\theta$ is a vector of unknown parameters.

Denoting the random effect size by $X$ and re-using the first part of the preceding calculation, the conditional density of an observation given significance and a particular sample size is

$$\begin{aligned}
\frac{d}{dt}P\{T_j \leq t | T_j > c_j, N_j = n_j\} &= \frac{d}{dt}\,\frac{P\{c_j < T_j \leq t | N_j = n_j\}}{P\{T_j > c_j | N_j = n_j\}} \\
&= \frac{d}{dt}\,\frac{\int_0^\infty P\{c_j < T_j \leq t | N_j = n_j, X = \texttt{es}\} g_\theta(\texttt{es}|n_j)\, d\texttt{es}}{\int_0^\infty P\{T_j > c_j | N_j = n_j, X = \texttt{es}\} g_\theta(\texttt{es}|n_j)\, d\texttt{es}} \\
&= \frac{d}{dt}\,\frac{\int_0^\infty \left[\texttt{p}(t, f_1(n_j) f_2(\texttt{es})) - \texttt{p}(c_j, f_1(n_j) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n_j)\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p}(c_j, f_1(n_j) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n_j)\, d\texttt{es}} \\
&= \frac{\int_0^\infty \frac{d}{dt}\texttt{p}(t, f_1(n_j) f_2(\texttt{es}))\, g_\theta(\texttt{es}|n_j)\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p}(c_j, f_1(n_j) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n_j)\, d\texttt{es}} \\
&= \frac{\int_0^\infty \texttt{d}(t, f_1(n_j) f_2(\texttt{es}))\, g_\theta(\texttt{es}|n_j)\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p}(c_j, f_1(n_j) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n_j)\, d\texttt{es}},
\end{aligned}$$

where moving the derivative through the integral sign is justified by dominated convergence. The likelihood function is a product of $k$ such terms. In the main paper, the simplifying assumption that sample size and effect size are independent before selection means that $g_\theta(\texttt{es}|n_j)$ is replaced by $g_\theta(\texttt{es})$, yielding Expression (3.2).

In the problem of estimating power under heterogeneity in effect size, the unknown parameters consist of the vector $\theta$ in the density of effect size. Let $\widehat{\theta}$ denote the maximum likelihood estimate of $\theta$. This yields a maximum likelihood estimate of the true power of each individual test in the sample, and then the estimates are averaged to obtain an estimate of mean power. We now give details.

Consider randomly sampling a single test from the population of tests that were significant the first time they were carried out. Temporarily changing the meaning of the symbol $T_j$, let $T_1$ denote the value of the test statistic the first time the hypothesis is tested, and let $T_2$ denote the value of the test statistic the second time the hypothesis is tested. Conditionally on fixed values of sample size $n$ and effect size $\texttt{es}$, $T_1$ and $T_2$ are independent. By Principle 1, population mean power after selection given the sample size $n$ is

$$P\{T_2 > c | T_1 > c\} = \sum_n P\{T_2 > c | T_1 > c, N = n\} P\{N = n | T_1 > c\} \tag{3.1}$$

This is the expression we seek to estimate. To obtain a useable formula for $P\{T_2 > c | T_1 > c, N = n\}$, we will first derive $g_\theta(\texttt{es}|n, T_1 > c)$, the conditional density of effect size given significance and a fixed sample size. Denote the random effect size by $X$ and continue to let $\texttt{es}$ represent a particular value of $X$. As before, assume that effect size has a continuous distribution with density $g_\theta(\texttt{es}|n)$. Then

$$\begin{aligned}
g_\theta(\texttt{es}|n, T_1 > c) &= \frac{d}{d\texttt{es}} P\{X \le \texttt{es}|T_1 > c, N = n\} \\
&= \frac{d}{d\texttt{es}} \frac{P\{X \le \texttt{es}, T_1 > c, N = n\}}{P\{T_1 > c, N = n\}} \\
&= \frac{d}{d\texttt{es}} \frac{\int_0^{\texttt{es}} P\{T_1 > c|X = y, N = n\} g_\theta(y|n) P\{N = n\} dy}{\int_0^\infty P\{T_1 > c|X = y, N = n\} g_\theta(y|n) P\{N = n\} dy} \\
&= \frac{\frac{d}{d\texttt{es}} \int_0^{\texttt{es}} P\{T_1 > c|X = y, N = n\} g_\theta(y|n) dy}{\int_0^\infty P\{T_1 > c|X = y, N = n\} g_\theta(y|n) dy} \\
&= \frac{P\{T_1 > c|X = \texttt{es}, N = n\} g_\theta(\texttt{es}|n)}{\int_0^\infty P\{T_1 > c|X = y, N = n\} g_\theta(y|n) dy} \\
&= \frac{\left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n)}{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(y))\right] g_\theta(y|n) dy}, \qquad (3.2)
\end{aligned}$$

an expression reminiscent of Principle 2, since the denominator is expected power given $N = n$. In the calculation that follows, Expression 3.2 will be substituted for $g_\theta(\texttt{es}|n, T_1 > c)$.

The true (not estimated, yet) probability that the test will be significant upon replication given the sample size is $P\{T_2 > c|T_1 > c, N = n\}$

$$\begin{aligned}
&= \frac{P\{T_1 > c, T_2 > c, N = n\}}{P\{T_1 > c, N = n\}} \\
&= \frac{\int_0^\infty P\{T_1 > c|T_2 > c, X = \texttt{es}, N = n\} g_\theta(\texttt{es}|n, T_1 > c) P\{T_1 > c, N = n\} d\texttt{es}}{P\{T_1 > c, N = n\}} \\
&= \int_0^\infty P\{T_1 > c|X = \texttt{es}, N = n\} g_\theta(\texttt{es}|n, T_1 > c) d\texttt{es} \\
&= \int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n, T_1 > c) d\texttt{es} \\
&= \int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] \frac{\left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n)}{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(y))\right] g_\theta(y|n) dy} d\texttt{es} \\
&= \frac{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right]^2 g_\theta(\texttt{es}|n)\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n)\, d\texttt{es}}. \qquad (3.3)
\end{aligned}$$

This expression could have been obtained with less effort by applying Principle 3 to the sub-population of tests based on a sample of size $n$. Consider it a cross-check.

Substituting (3.3) into (3.1) yields

$$P\{T_2 > c|T_1 > c\} = \sum_n \frac{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right]^2 g_\theta(\texttt{es}|n)\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))\right] g_\theta(\texttt{es}|n)\, d\texttt{es}} P\{N = n|T_1 > c\}. \qquad (3.4)$$

Expression 3.4 has two unknown quantities, the parameter $\theta$ of the effect size distribution, and $P\{N = n|T_1 > c\}$. For the former quantity, we use the maximum likelihood estimate, while the $P\{N = n|T_1 > c\}$ values are estimated by the empirical relative frequencies of sample size (which is the non-parametric maximum likelihood estimate). The result is a maximum likelihood estimate of population power given significance:

$$\frac{1}{k}\sum_{j=1}^{k}\frac{\int_0^\infty \left[1-\texttt{p}(c_j\texttt{,}f_1(n_j)f_2(\texttt{es}))\right]^2 g_{\widehat{\theta}}(\texttt{es}|n_j)\,d\texttt{es}}{\int_0^\infty \left[1-\texttt{p}(c_j\texttt{,}f_1(n_j)f_2(\texttt{es}))\right] g_{\widehat{\theta}}(\texttt{es}|n_j)\,d\texttt{es}}\,.$$

In the simulations, the density $g$ of effect size is gamma, there is no dependence on $n$, and the parameter $\theta$ is the pair $(a, b)$ that parameterize the gamma distribution.

# 4 Simulation

## 4.1 Direct simulation from the distribution of the test statistic given significance

To study the behaviour of an estimation method under selection for significance, it is natural to simulate test statistics from the distribution that applies before selection, and then discard the ones that are not significant. But if one can simulate from the joint distribution of sample size and effect size after selection, the wasteful discarding of non-significant test statstics can be avoided. The idea is to do the simulation in two stages. First, simulate pairs from the joint distribution of sample size and effect size after selection, and calculate a non-centrality parameter using Expression (ncpmult). Then using that `ncp` value, simulate from the distribution of the test statistic given significance. We will now show how to do the second step.

It is well known that if $F(t)$ is a cumulative distribution function of a continuous random variable and $Y$ is uniformly distributed on the interval from zero to one, then the random variable $T = F^{-1}(Y)$ has cumulative distribution function $F(t)$. In this case the cumulative distribution function from which we wish to simulate is

$$\begin{aligned}
P\{T \le t|T > c, X = \texttt{es}, N = n\} &= \frac{P\{T \le t, T > c|X = \texttt{es}, N = n\}}{P\{T > c|X = \texttt{es}, N = n\}} \\
&= \frac{P\{c < T \le t|X = \texttt{es}, N = n\}}{P\{T > c|X = \texttt{es}, N = n\}} \\
&= \frac{\texttt{p(}t\texttt{,ncp)} - \texttt{p(}c\texttt{,ncp)}}{1 - \texttt{p(}c\texttt{,ncp)}}
\end{aligned}$$

for $t > c$, where as usual $\texttt{ncp} = f_1(n)f_2(\texttt{es})$. To obtain the inverse, set $y$ equal to the probability and solve for $t$, as follows.

$$\begin{aligned}
& y = \frac{\texttt{p(}t\texttt{,ncp)} - \texttt{p(}c\texttt{,ncp)}}{1 - \texttt{p(}c\texttt{,ncp)}} \\
\Leftrightarrow\ & y\,(1 - \texttt{p(}c\texttt{,ncp)}) = \texttt{p(}t\texttt{,ncp)} - \texttt{p(}c\texttt{,ncp)} \\
\Leftrightarrow\ & \texttt{p(}t\texttt{,ncp)} = y\,(1 - \texttt{p(}c\texttt{,ncp)}) + \texttt{p(}c\texttt{,ncp)} = \gamma y + 1 - \gamma \\
\Leftrightarrow\ & t = \texttt{q(}\gamma y + 1 - \gamma\texttt{,ncp)},
\end{aligned}$$

where $\gamma = 1 - \texttt{p(}c\texttt{,ncp)}$ is the power of the test. Accordingly, let $Y$ be a Uniform (0,1) random variable. The significant test statistic is

$$\begin{aligned} T &= \texttt{q(}\gamma Y + 1 - \gamma\texttt{,ncp)} \\ &= \texttt{q(}1 + \gamma(Y-1)\texttt{,ncp)} \\ &= \texttt{q(}1 - \gamma(1-Y)\texttt{,ncp)}\,. \end{aligned}$$

Since $1-Y$ also has a Uniform (0,1) distribution, one may proceed as follows. First calculate the non-centrality parameter $\texttt{ncp} = f_1(n)f_2(\texttt{es})$, and the power value $\gamma = 1 - \texttt{p(}c\texttt{,ncp)}$. Then calculate the significant test statistic

$$T = \texttt{q(}1 - \gamma U\texttt{,ncp)}\,, \tag{4.1}$$

where $U$ is a pseudo-random variate from a Uniform (0,1) distribution. In R, the process can be applied to a vector of `ncp` values and a vector of independent $U$ values of the same length.

Again, this is the second step. The first step is to simulate a collection of `ncp` values using the desired joint distribution of sample size and effect size after selection for significance. Naturally, simulation is is easiest if sample size and effect size come from well-known distributions with built-in random number generation, and if sample size and effect size are specified to be independent after selection. In one of our simulations, sample size and effect size after selection were correlated. The next section describes how this was done.

## 4.2 Correlated sample size and effect size

Let effect size $X$ have density $g_\theta(\texttt{es})$; conditionally on $X = \texttt{es}$, let the distribution of sample sizesample size be Poisson distributed with expected value $\exp(\beta_0 + \beta_1\texttt{es})$. This is standard Poisson regression. Simulation from the joint distribution is easy. One simply simulates an effect size `es` according to the density $g$, computes the Poisson parameter $\lambda = \exp(\beta_0+\beta_1\texttt{es})$, and then samples a value $n$ from a Poisson distribution with parameter $\lambda$. The challenge is to choose the parameters $\theta, \beta_0$ and $\beta_1$ so that (a) the population mean power has a desired value, and at the same time (b) the population correlation between sample size and effect size has a desired value. Population mean power is

$$\gamma = \int_0^\infty \sum_n \left[1 - \texttt{p(}c\texttt{,}f_1(n)f_2(\texttt{es})\texttt{)}\right] P\{N = n | X = \texttt{es}\} g_\theta(\texttt{es}) d\texttt{es}\,.$$

Given values of $\theta, \beta_0$ and $\beta_1$, this expression can be calculated by numerical integration; recall that $P\{N = n | X = \texttt{es}\}$ is a Poisson probability.

The population correlation between sample size and effect size is

$$\rho = \frac{E(XN) - E(X)E(N)}{SD(X)SD(N)}\,,$$

where $SD(\cdot)$ refers to the population standard deviation of something. The quantities $E(X)$ and $SD(X)$ are direct functions of $\theta$. The standard deviation of sample size $SD(N) = \sqrt{E(N^2) - [E(N)]^2}$, where

$$\begin{aligned} E(N) &= E(E[N|X]) \\ &= \int_0^\infty E[N|X = \texttt{es}]\, g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty e^{\beta_0 + \beta_1 \texttt{es}} g_\theta(\texttt{es}) d\texttt{es} \end{aligned}$$

and

$$\begin{aligned} E(N^2) &= E(E[N^2|X]) \\ &= E(Var(N) + E(N)^2|X) \\ &= \int_0^\infty \left(e^{\beta_0+\beta_1\texttt{es}} + e^{2\beta_0+2\beta_1\texttt{es}}\right) g_\theta(\texttt{es}) d\texttt{es}\,. \end{aligned}$$

Finally,

$$\begin{aligned} E(XN) &= \int_0^\infty \sum_n \texttt{es}\, n\, P\{N = n|X = \texttt{es}\} g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty \texttt{es}\, E(N|X = \texttt{es}) g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty \texttt{es}\, e^{\beta_0+\beta_1\texttt{es}} g_\theta(\texttt{es}) d\texttt{es}\,. \end{aligned}$$

All these expected values can be calculated by numerical integration using R's `integrate` function, so that the correlation $\rho$ can be evaluated for any set of $\theta, \beta_0$ and $\beta_1$ values.

In our simulation of correlated sample size and effect size, $g_\theta(\texttt{es})$ was a beta density, re-parameterized so that $\theta = (\mu, \sigma^2)$ consisted of the mean $\mu$ and variance $\sigma^2$. Conditionally on effect size, sample size was Poisson distributed with expected value $\exp(\beta_0 + \beta_1\texttt{es})$. We set the variance of effect size $\sigma^2$ to a fixed value of 0.09, so that the standard deviation of effect size after selection was 0.30, a high value. Given any mean effect size $\mu$ and slope $\beta_1$, the parameter $\beta_0$ (the intercept of the Poisson regression) was adjusted so that expected sample size at the mean value was equal to 86: $\beta_0 = \ln(86) - \beta_1\mu$.

With these constraints, the population mean power $\gamma$ and correlation $\rho$ were a function of the two free parameters $\mu$ and $\beta_1$. Let $\gamma_0$ be a desired value of mean power; for example, $\gamma_0 = 0.5$. Let $\rho_0$ be a desired value of the correlation between sample size and effect size; for example, $\rho_0 = -0.8$. Values of $\mu$ and $\beta_1$ were locating by numerically minimizing the function $f(\mu, beta_1) = |\gamma - \gamma_0| + |\rho - \rho_0|$. We used R's optim function.

## 4.3 How selection affects the joint distribution of $N$ and es

Sometimes it is desirable to specify the distributions of sample size and effect size before selection for significance. In this case we found it necessary to simulate test statistics before selection and then literally discard non-significant results. In principle, Expression (4.1) could be used to generate significant test statistics provided that one could simulate $(n, \texttt{es})$ pairs from the joint distribution that obtains after selection. The following Principle (not given in the main paper) shows why this is difficult. The similarity to Principle 2 is remarkable.

**Principle 6** *The effect of selection for significance is to multiply the joint distribution of sample size and effect size by power for that sample size and effect size, divided by population mean power before selection.*

**Proof** Note that power for a given sample size and effect size is $P\{T > c|X = \texttt{es}, N = n\}$. Suppose effect size is discrete. Then

$$\begin{aligned} P\{X = \texttt{es}, N = n|T > c\} &= \frac{P\{X = \texttt{es}, N = n, T > c\}}{P\{T > c\}} \\ &= \frac{P\{T > c|X = \texttt{es}, N = n\}P\{X = \texttt{es}, N = n\}}{E(G)} \\ &= \left(\frac{P\{T > c|X = \texttt{es}, N = n\}}{E(G)}\right) P\{X = \texttt{es}, N = n\}\,, \end{aligned}$$

where as in Section 2, $E(G)$ is expected power before selection, equal to $P\{T > c\}$ by Principle 1.

Suppose that effect size is continuous with density $g(\texttt{es})$. The joint distribution of sample size and effect size before selection is determined by $P\{N = n|X = \texttt{es}\}g(\texttt{es})$. The joint distribution after selection is determined by

$$\begin{aligned} P\{N = n|X = \texttt{es}, T > c\}g(\texttt{es}|T > c) &= \frac{P\{T > c|X = \texttt{es}, N = n\}P\{N = n|X = \texttt{es}\}g(\texttt{es})}{g(\texttt{es}|T > c)P\{T > c\}} g(\texttt{es}|T > c) \\ &= \left(\frac{P\{T > c|X = \texttt{es}, N = n\}}{E(G)}\right) P\{N = n|X = \texttt{es}\}g(\texttt{es})\,. \end{aligned}$$

It is also possible to write the joint distribution of sample size and effect size as the conditional density of effect size given sample size, times the discrete probability of sample size. That is, the joint distribution before selection is determined by $g(\texttt{es}|N = n)P\{N = n\}$, and the joint distribution after selection is determined by

$$\begin{aligned} g(\texttt{es}|N = n, T > c)P\{N = n|T > c\} &= \frac{d}{d\texttt{es}} P\{X \leq \texttt{es}|N = n, T > c\}P\{N = n|T > c\} \\ &= \frac{d}{d\texttt{es}} \frac{P\{X \leq \texttt{es}, N = n, T > c\}}{P\{N = n, T > c\}} \frac{P\{N = n, T > c\}}{P\{T > c\}} \\ &= \frac{1}{E(G)} \frac{d}{d\texttt{es}} \int_0^{\texttt{es}} P\{T > c|X = y, N = n\}g(y|N = n)P\{N = n\}\,dy \\ &= \frac{P\{T > c|X = \texttt{es}, N = n\}g(\texttt{es}|N = n)P\{N = n\}}{E(G)} \\ &= \left(\frac{P\{T > c|X = \texttt{es}, N = n\}}{E(G)}\right) g(\texttt{es}|N = n)P\{N = n\} \quad \blacksquare \end{aligned}$$

In terms of simulation, these formulas show that if the joint distribution of effect size and sample size before selection is familiar and convenient, the distribution after selection will be unfamiliar and inconvenient, because sample size and effect size both appear in the expression for power: $P\{T > c|X = \texttt{es}, N = n\} = 1 - \texttt{p(}c\texttt{,}f_1(n)f_2(\texttt{es})\texttt{)}$. In this case the most direct way to sample exactly from the distribution of $N$ and `es` after selection (and for us, the only way) is to literally select values corresponding to significant results.

# References

[1] Agresti, A. (1990), *Categorical data analysis.* Hoboken, N.J.: Wiley.

[2] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd Edition), Hilsdale, New Jersey: Erlbaum.

[3] Stuart, A. and Ord, J. K. (1999). *Kendall's Advanced Theory of Statistics, Vol. 2: Classical Inference & the Linear Model* (5th ed.). New York: Oxford University Press.